FabCafe
what do you fab?
bangkok

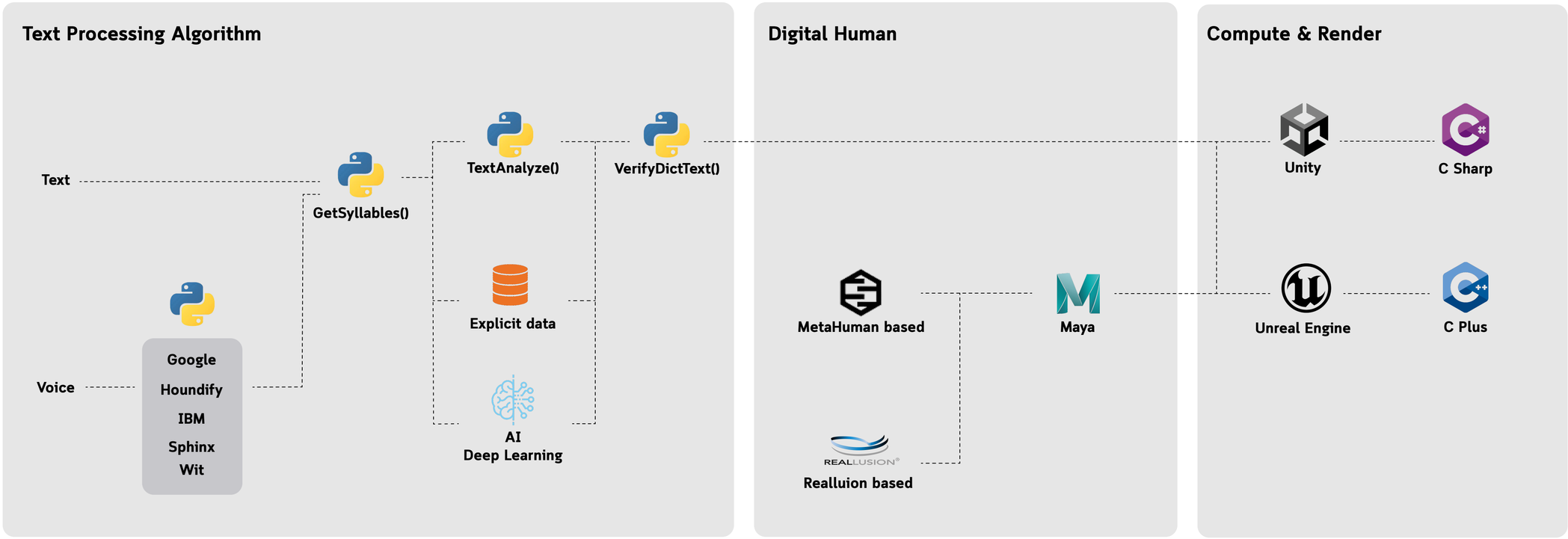
Speakable
Digital Human

Contents

Process

FabCafe what do you fab? bangkok

Source	Speech to Text	Chunk	Analyze	Verify	Human Model	Blend Shape	Computing Engine	Animate



รับข้อมูลเข้า โดยมีสองกรณี คือ สายอักขระ และ เสียงพูด

ข้อมูลนำเข้าที่เป็นคลิปเสียงพูด จะต้องนำมาผ่านกระบวนการ Speech to Text เพื่อแปลงให้เป็นสายอักขระ ส่วนข้อมูลที่เป็นสายอักขระอยู่แล้ว ไม่จำเป็นต้องผ่านขั้นตอนนี้

Chunking คือการแตกข้อมูลออกเป็นส่วนย่อย ๆ ในที่นี้คือการแตกสายอักขระออกเป็นOrdered List ของแต่ละพยางค์

นำแต่ละพยางค์ไปประมวลผล เพื่อหา พยัญชนะต้น, สระ, และ ตัวสะกด ตามหลักไวยากรณ์ภาษาไทย แบ่งออกเป็น3วิธี
1.ใช้ algorithm ในการแยกส่วนของพยางค์
2.ใช้ Dataset ที่ระบุทุกคำไว้แล้ว
3.ใช้ Dataset ที่ระบุคำยากไว้ และใช้ algorithm ในคำง่าย

ตรวจสอบว่าพยางค์ใด มีส่วนประกอบที่ขัดกับหลักภาษาไทยหรือไม่

โมเดลมนุษย์3มิติ ที่จะนำมาใช้ในการพูด

สำหรับโมเดลที่เป็น MetaHuman based ใบหน้าจะถูกขยับด้วยระบบ Rig หรือ Bones ซึ่งจะต้องนำไปเปลี่ยนให้เป็นระบบ Blend Shape ด้วยโปรแกรม Maya
ส่วนโมเดลที่เป็น Reallusion based จะใช้ระบบ Blend Shape อยู่แล้ว

นำสองส่วน ได้แก่ โมเดล และ algorithm มารวมกันใน Computing Engine เช่น Unity และ Unreal Engine

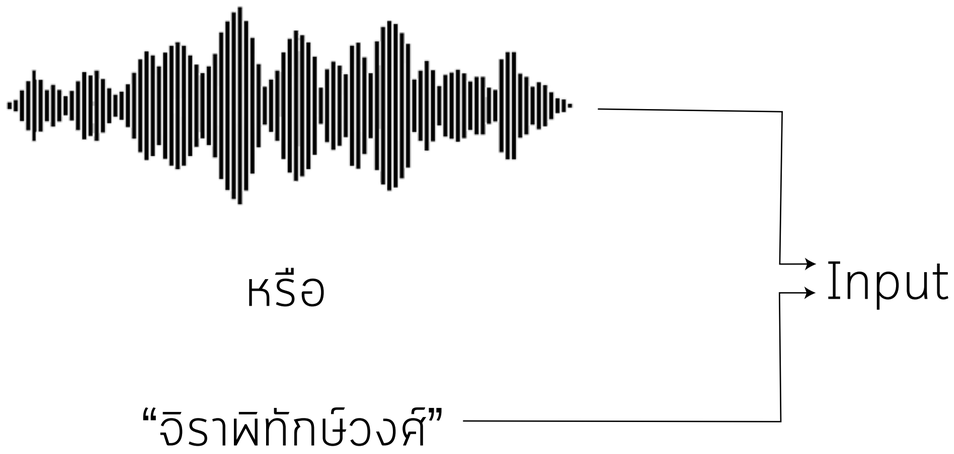
ใช้สคริปต์ในการ เรียกใช้ algorithm และควบคุม Blend Shape เพื่อขยับปาก

Possible Path

Flow

Source

Process 01



Source คือการใส่ Input ให้กับ Text Processing Algorithm โดย Input อาจมาจาก การพิม การพูดของมนุษย์ หรือจะเป็นข้อความที่ถูกสร้างขึ้นด้วย AI ก็ได้

Source มีสองประเภท คือ

1. **ข้อความ** โดยอยู่ในรูปของ String หรือ สายอักขระ
2. **เสียงพูด** มีสองประเภทย่อย คือ
   - 2.1 คลิปเสียงพูด ซึ่งอาจเป็นเสียงคน หรือเสียงสังเคราะห์ก็ได้
   - 2.2 เสียงพูดสด จากไมค์

Speech to Text

Process 02

Speech to Text คือเทคโนเลยีที่ใช้ AI ในการแปลงเสียงพูด เป็น String หรือสายอักขระ สำหรับใช้ในการประมวลผลต่อไป

ปัจจุบันมีหลายบริษัทที่เปิด Open Source และ API สำหรับเทคโนโลยี Speech to Text ได้แก่

1. Google and GoogleCloud
2. Houdify by SoundHoud
3. IBM
4. Sphinx
5. Wit

เมื่อจบขั้นตอนนี้ จะได้ข้อมูลในรูป String

Speech to Text API

Google
CMUSphinx
SoundHound
IBM
wit.ai

จิราพิทักษ์วงศ์

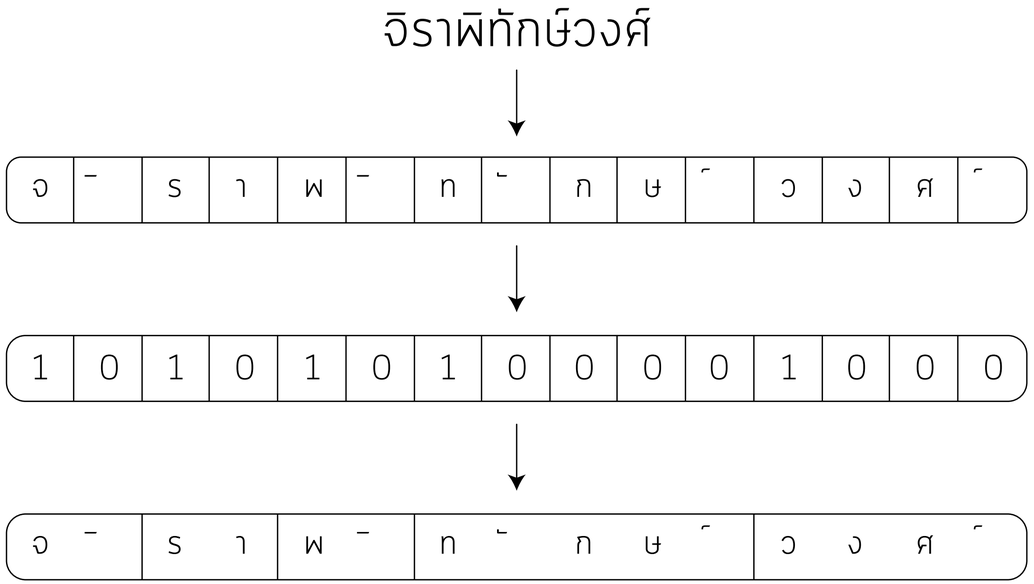
Chunk

Process 03

Chunking คือการแตกข้อมูลออกเป็นส่วนย่อย ๆ ในที่นี้ คือการแตกสายอักขระออกเป็นOrdered List ของแต่ละพยางค์

การทำ Syllables Chunking จะเป็นการนำ ตัวอักษรทุกตัว จากสายอักขระ มากางลงบน Array (ภาพบน) จากนั้นใช้ Deep Learning ควบคู่ไปกับ Explicit Algorithm เพื่อแปลงให้เป็น Digital Array โดย
1 หมายถึง น่าจะเปนตัวอักษรแรกของพยางค์
0 หมายถึง น่าจะไม่ใช่ตัวอักษรแรกของพยางค์
จากนั้นจึงทำการตัดสตริงตาม Index ของ Digital Array ที่เป็น 1

เมื่อจบขั้นตอนนี้ จะได้ข้อมูลในรูป List ของ String



Analyze

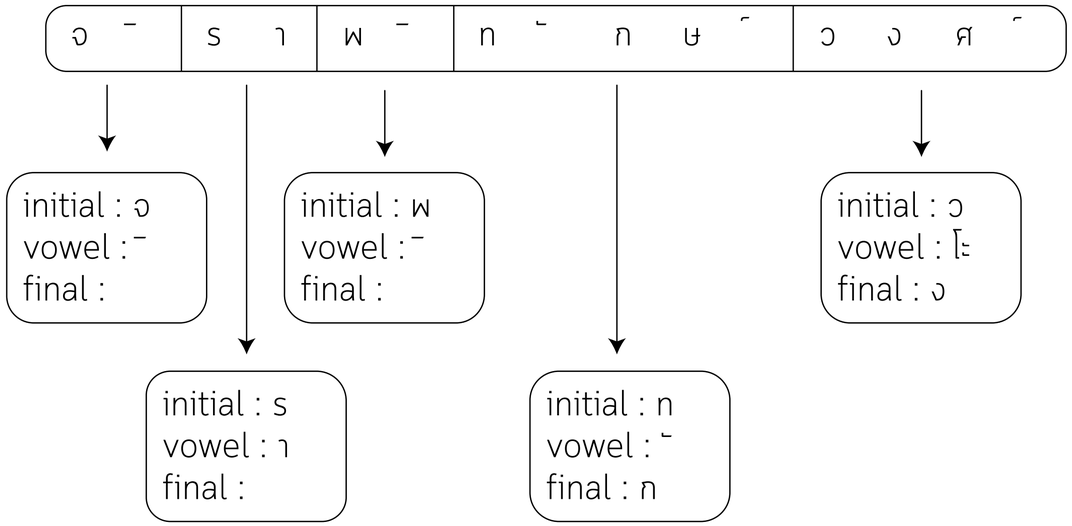
Process 04

Analyze คือการแตกส่วนประกอบของแต่ละพยางค์ให้เป็น พยัญชนะต้น, สระ, และตัวสะกด

การ Analyze ทำได้ 3 วิธี คือ

1. ใช้ **AI-DeepLearning**
   คือการเทรน AI เพื่อจำแนกส่วนประกอบโดยเฉพาะ
2. ใช้ **Explicit Algorithm**
   คือการเขียนAlgorithmที่ถอดไวยากรณ์ภาษาไทย โดยกฎทั้งหมดจะถูกเขียนไว้ชัดเจน ไม่มีการสุ่ม และไม่มีการใช้ AI เข้ามาเกี่ยวข้อง
3. ใช้ **Explicit Data**
   คือการสร้าง Dataframe ที่ระบุคำทุกคำในภาษาไทย รวมถึงบอกว่าแต่ละคำมีโครงสร้างส่วนประกอบอย่างไร
4. ใช้ **ข้อ1 และข้อ2 ร่วมกัน**

เมื่อจบขั้นตอนนี้ จะได้ List ของ Map ของแต่ละพยางค์ซึ่งระบุส่วนประกอบเอาไว้



Verify

Process 05

Verify คือการตรวจสอบ ว่าในแต่ละพยางค์ ที่ผ่านการ Analyze มา มีคำใดที่ขัดกับหลักภาษาไทย และ หลักการออกเสียงหรือไม่หรือไม่ เพื่อนำพยางค์ที่ผิดไปแก้ไข หรือ ตัดทิ้ง

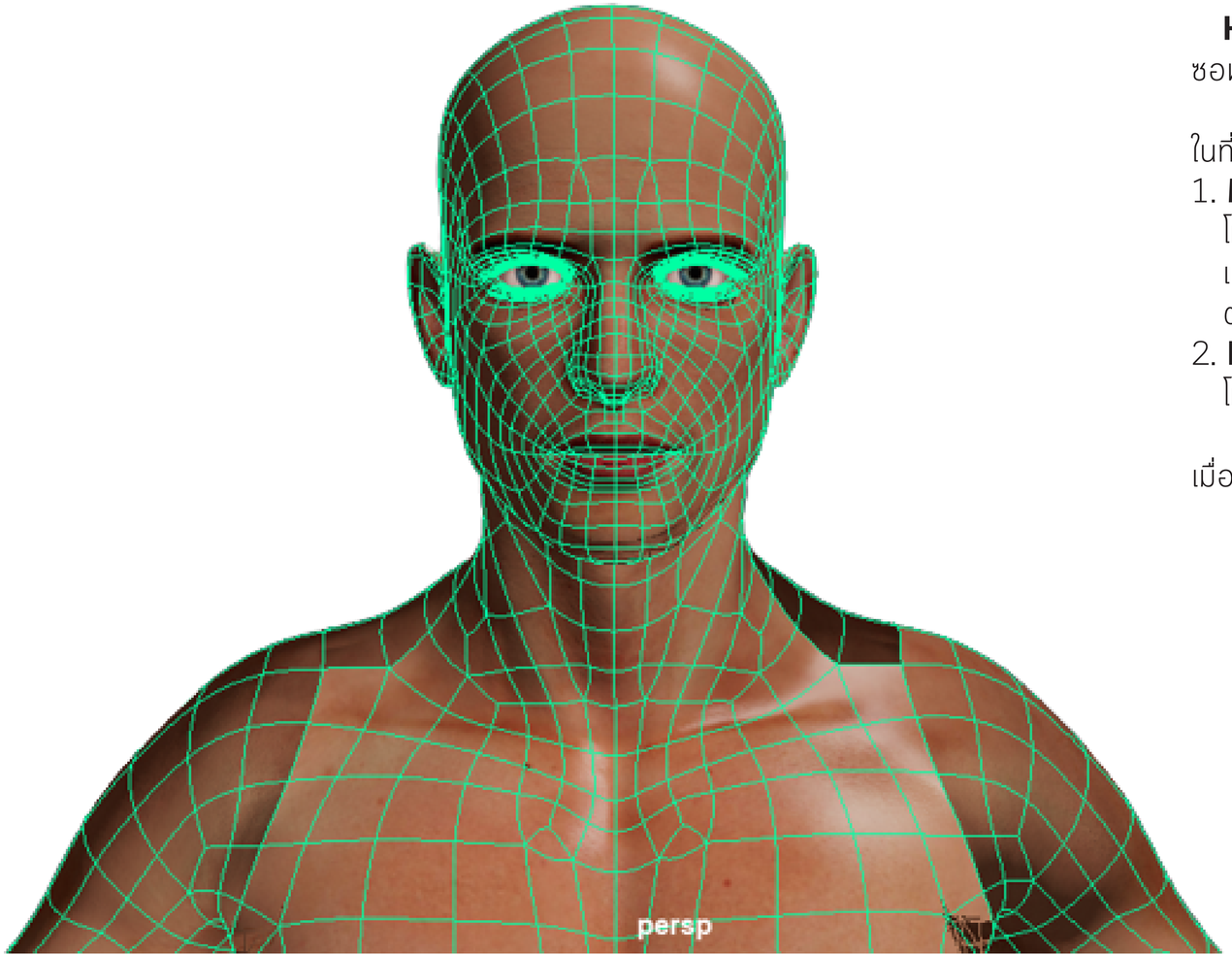
“กฮ”

“ไสว”

“สวัส”

Human Model

Process 06



Human Model หมายถึงไฟล์ 3มิติ รูปมนุษย์ ซึ่งปั้นจากซอฟแวร์ใดก็ได้ มีRigหรือไม่ก็ได้ แต่จะต้องสามารถอ้าปากได้

ในที่นี้ จะกล่าวถึงโมเดลที่ปั้นจากโมเดลพื้นฐาน แบ่งเป็น

1. **MetaHuman based**
   โมเดลที่ได้ จะมีการ Rig ใบหน้า ด้วยระบบ Bones และจะต้องเปลี่ยนไปเป็น ระบบ BlendShape ในขั้นตอนต่อไป
2. **Reallusion CC based**
   โมเดลที่ได้ จะมีการ Rig ใบหน้า ด้วยระบบ BlendShape

เมื่อจบขั้นตอนนี้จะได้ไฟล์สกุล FBX

Blend Shape

Process 07

Blend Shape คือชื่อเรียกในโปรแกรม Maya หมายถึง การเคลื่อนย้าย Vertex ของโมเดลตัวหนึ่ง ให้มีรูปทรงเหมือนโมเดลอีกตัวหนึ่ง

ในขั้นตอนนี้ จะเป็น Duplicate ส่วนหัว ของตัวตั้งต้นออกมาจำนวนมาก เพื่อนำมาขยับปากเป็นทรงต่างๆที่แตกต่างกัน หลังจากนั้น นำโมเดลตั้งต้น มัน Blend Shape กับโมเดลที่ Duplicate ออกมา ซึ่งจะทำให้เรามารถควบคุมทรงปากของโมเดลตั้งต้นได้ และสามารถExport เพื่อนำไปควบคุมในโปรแกรมอื่นๆได้

เมื่อจบขั้นตอนนี้จะได้ไฟล์สกุล FBX

Computing Engine

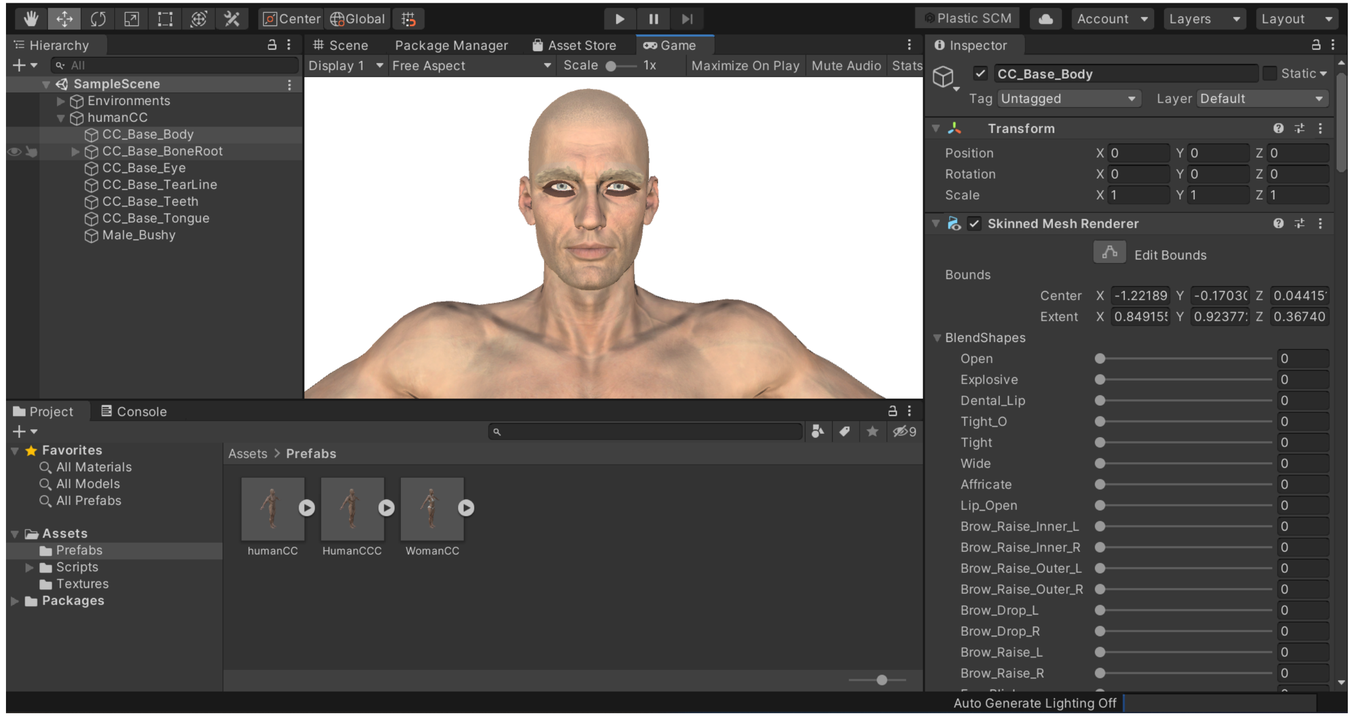
Process 08

Computing Engine หมายถึงซอฟท์แวร์ที่ทำหน้าที่ในการ ประมวลผล และเรนเด้อ ซึ่งจะสามารถเขียนสคริปต์เพื่อควบคุมการทำงานของสิ่งต่างๆ ภายใน Scene ได้

ในขั้นตอนนี้ จะเป็นการนำ Human Model และ ALgorithm มาไว้รวมกัน

Engine ที่ใช้ได้มีสองตัว คือ
1. **Unity**
2. **Unreal Engine**

เมื่อจบขั้นตอนนี้ จะได้ไฟล์โปรเจค Unity และ/หรือ Unreal Engine



Animate

Process 09

initial : ก
vowel : ั
final : ก

Animate หรือการขยับ เป็นขั้นตอนใหญ่ ซึ่งประกอบด้วย

1. Map ข้อมูลจาก Text Processing Algorithm เข้ากับรูปปากแบบต่างๆ
2. เรียกใช้ Algorithm ด้วยสคริปต์ ซึ่งแบ่งเป็น C# สำหรับ Unity และ C++ สำหรับ Unreal Engine
3. รับ Input ตามจุดประสงค์ เช่น จากผู้ใช้ จากAI เป็นต้น
4. จัดกล้อง แสง และสภาพแวดล้อม เพื่อนRender Animation ออกมา

ตอนนี้สามารถพัฒนาไปได้เรื่อยๆไม่สิ้นสุด ขึ้นอยู่กับจุดะสงค์การใช้งาน ความแม่นยำที่ต้องการ ความหลายหลายงแต่ละภาษา

งจบขั้นตอนนี้ จะได้ไฟล์โปรเจค ซึ่งภายในโปรเจคจะมีตัวจรซึ่งสามารถขยับปากตามสิ่งที่พูดได้

Show Case

Text Processing Algorithm

FabCafe bangkok
what do you fab?

Show Case

Computing Engine

Future Improvement

Timing

คือการควบคุมจังหวะการขยับปากให้ตรงกับเสียงพูด สามารถทำได้สองวิธี คือ

1. ในขั้นตอน Speech to text ใช้ Machine Learning Chunk เสียงพูดออกเป็นพยางค์ตั้งแต่แรก แล้วเก็บข้อมูลเวลาของแต่ละพยางค์มาใช้ในการ Animate
2. หลังจากขั้นตอน Verify นำ ข้อมูลแต่ละพยางค์มา Mapกับไฟล์เสียง ด้วย Machine Learningพยางค์ต่อพยางค์ แล้วสร้างข้อมูลเวลาขึ้นมา

Accuracy

คือการเพิ่มความสมจริงในการขยับปากตามคำพูดแต่ละคำ ซึ่งจำได้ก็ต่อเมื่อมี Timing ที่ถูกต้อง

Transition

คือ การเชื่อมต่อAnimationของแต่ละพยางค์ เพื่อให้สมจริงยิ่งขึ้น